

# ANALYSIS OF CONVERGENCE RATES OF SOME GIBBS SAMPLERS ON CONTINUOUS STATE SPACES

AARON SMITH

## 1. ABSTRACT

We use a non-Markovian coupling and small modifications of techniques from the theory of finite Markov chains to analyze some Markov chains on continuous state spaces. The first is a generalization of a sampler introduced by Randall and Winkler, the second a Gibbs sampler on narrow contingency tables.

## 2. INTRODUCTION

The problem of sampling from a given distribution on high-dimensional continuous spaces arises in the computational sciences and Bayesian statistics, and a frequently-used solution is Markov chain Monte Carlo (MCMC); see [16] for many examples. Because MCMC methods produce good samples only after a lengthy mixing period, a long-standing mathematical question is to analyze the mixing times of the MCMC algorithms which are in common use. Although there are many mixing conditions, the most commonly used is called the mixing time, and is based on the total variation distance:

For measures  $\nu, \mu$  with common measurable  $\sigma$ -algebra  $\mathcal{A}$ , the *total variation distance* between  $\mu$  and  $\nu$  is

$$\|\mu - \nu\|_{TV} = \sup_{A \in \mathcal{A}} (\mu(A) - \nu(A))$$

For an ergodic discrete-time Markov chain  $X_t$  with unique stationary distribution  $\pi$ , the *mixing time* is

$$\tau(\epsilon) = \inf\{t : \|\mathcal{L}(X_t) - \pi\|_{TV} < \epsilon\}$$

Although most scientific and statistical uses of MCMC methods occur in continuous state spaces, much of the mathematical mixing analysis has been in the discrete setting. The methods that have been developed for discrete chains often break down when used to analyze continuous chains, though there are some efforts, such as [28] [24] [18], to create general techniques. This paper extends the author's previous work in [26] and work of Randall and Winkler [22], and attempts to provide some more

---

*Date:* June 13, 2012.

examples of relatively sharp analyses of continuous chains similar to those used to develop the discrete theory.

The first process that we analyze is a Gibbs sampler on the simplex with a very restricted set of allowed moves. Fix a finite group  $G$  of size  $n$  with symmetric generating set  $R$  of size  $m$ . For unity of notation, label the group elements with the integers from 1 to  $n$ . We consider the process  $X_t[g]$  on the simplex  $\Delta_G = \{X \in \mathbb{R}^n \mid \sum_{g \in G} X[g] = 1; X[g] \geq 0\}$ . At each step, choose  $g \in G$ ,  $r \in R$  and  $\lambda \in [0, 1]$  uniformly, and set

$$(1) \quad \begin{aligned} X_{t+1}[g] &= \lambda(X_t[g] + X_t[gr]) \\ X_{t+1}[gr] &= (1 - \lambda)(X_t[g] + X_t[gr]) \end{aligned}$$

For all other  $h \in G$  set  $X_{t+1}[h] = X_t[h]$ . Let  $U_G$  be the uniform distribution on  $\Delta_G$ ; this is also the stationary distribution of  $X_t$ . Also consider a random walk  $Z_t$  on  $G$ , where in each stage we choose  $g \in G$  and  $r \in R$  uniformly at random and set  $Z_{t+1} = gr$  if  $Z_t = g$ , set  $Z_{t+1} = g$  if  $Z_t = gr$ , and  $Z_{t+1} = Z_t$  otherwise. This is the standard simple random walk on the Cayley graph, slowed down by a factor of about  $n$ . Let  $\hat{\gamma}$  be the spectral gap of the walk  $Z_t$ , and follow the notation that  $\mathcal{L}(X)$  denotes the distribution of a random variable  $X$ .

**Theorem 1** (Convergence Rate for Gibbs Sampler with Geometry). *For  $T > \frac{56k+178}{\hat{\gamma}} \log(n)$ ,*

$$\|\mathcal{L}(X_T) - U_G\|_{TV} \leq 14n^{-k}$$

*and conversely for  $T < \frac{k}{\hat{\gamma}}$ ,*

$$\|\mathcal{L}(X_T) - U_G\|_{TV} \geq \frac{1}{2}e^{-k} - 3n^{-\frac{1}{3}}$$

This substantially generalizes [22] and [26], from samplers corresponding to  $G = \mathbb{Z}_n$ , and  $R = \{1, -1\}$  or  $R = \mathbb{Z}_n \setminus \{0\}$  respectively, to general Cayley graphs. In addition to being of mathematical interest, this process is an example of a gossip process with some geometry, studied by electrical engineers and sociologists interested in how information propagates through networks; see [25] for a survey.

The proof of the upper bound will use an auxilliary chain similar to that found in [22], a coupling argument improved from [26], and an unusual use of comparison theory from [9]. The proof of the lower bound is elementary.

The next example consists of narrow contingency tables. Beginning with the work of Diaconis and Efron [6] on independence tests, there has been interest in finding efficient ways to sample uniformly from the collection of integer-valued matrices with given row and column sums. A great deal of this effort has been based on Markov chain Monte Carlo methods. While some of the efforts have dealt directly with Markov chains on these integer-valued matrices, much recent success, including [11] [21], has

involved using knowledge of Gibbs samplers on convex sets in  $\mathbb{R}^n$  and clever ways to project from the continuous chain to the desired matrices [20].

Unfortunately, while the general bounds are polynomial in the number of entries in the desired matrix, they often have a large degree and leading coefficient; see [17]. In this paper, we find some better bounds for very specific cases. Like the paper [26], this is part of an attempt to make further use of non-Markovian coupling techniques [13] [2] [5] [19] and also to expand the small set of carefully analyzed Gibbs samplers [22] [23] [7] [8]. In this case, the new techniques are two slight modifications of the path-coupling method introduced in [4]. In many path-coupling arguments, a burn-in argument is used to show that for most pairs of points in a metric space, there is a path along which the Markov transition kernel is contractive acting on any pair of points along the path. In this argument, we show that for all paths, the kernel is contractive acting on most pairs of points along the path. This type of modification seems likely to be useful only on continuous spaces.

We consider the following Gibbs sampler  $X_t[i, j]$  on the space  $M_n = \{X \in \mathbb{R}^{2n} : \sum_{i=1}^n X[i, j] = n \ \forall 1 \leq j \leq 2, \sum_{j=1}^2 X[i, j] = 2 \ \forall 1 \leq i \leq n, X[i, j] \geq 0\}$  of nonnegative  $n$  by 2 matrices with column sums fixed to be  $n$  and row sums fixed to be 2. To make a step of the Gibbs sampler, choose two distinct integers  $1 \leq i < j \leq n$  and update the four entries  $X_{t+1}[i, 1]$ ,  $X_{t+1}[i, 2]$ ,  $X_{t+1}[j, 1]$  and  $X_{t+1}[j, 2]$  to be uniform conditional on all other entries of  $X_t$ . Let  $U_n$  be the uniform distribution on  $M_n$  inherited from Lebesgue measure. Then we find the following reasonable bound on the mixing time of this sampler:

**Theorem 2** (Convergence Rate for Narrow Matrices). *For  $T > (31k + 81)n \log(n)$ ,*

$$\|\mathcal{L}(X_T) - U_n\|_{TV} \leq 13n^{-k}$$

*and conversely for  $T < (1 - k)n \log(n)$ , and  $n$  sufficiently large,*

$$\|\mathcal{L}(X_T) - U_n\|_{TV} \geq 1 - 2n^{-k}$$

### 3. GENERAL STRATEGY AND THE PARTITION PROCESS

Both of our bounds will be obtained using a similar strategy, ultimately built on the classical coupling lemma. We recall that a coupling of Markov chains with transition kernel  $K$  is a process  $(X_t, Y_t)$  so that marginally both  $X_t$  and  $Y_t$  are Markov chains with transition kernel  $K$ . Although we always couple entire paths  $\{X_t\}_{t=0}^T$  and  $\{Y_t\}_{t=0}^T$ , we often use the shorthand notation of saying that we are coupling  $X_t$  and  $Y_t$ . In order to describe a coupling, note that for both walks being studied, the evolution of the Markov chain  $X_t$  can be represented by  $X_{t+1} = f(X_t, i(t), j(t), \lambda(t))$ , where  $f$  is a deterministic function,  $i(t), j(t)$  are random coordinates (either elements of  $[n]$  or of a group  $G$ ), and  $\lambda(t)$  is drawn from Lebesgue measure on  $[0, 1]$ . These representations are given in equations (8) and (1) respectively. To couple  $X_t$  and  $Y_t$ , it is thus enough

to couple the update variables  $i(t)^\alpha, j(t)^\alpha, \lambda(t)^\alpha$ , with  $\alpha \in \{x, y\}$ , used to construct  $X_t$  and  $Y_t$  respectively.

Our couplings will provide bounds on mixing times through the following lemma (see [15], Theorem 5.2 - they work in discrete space, but their proof doesn't rely on this assumption):

**Lemma 3** (Fundamental Coupling Lemma). *If  $(X_t, Y_t)$  is a coupling of Markov chains,  $Y_0$  is distributed according to the stationary distribution of  $K$ , and  $\tau$  is the first time at which  $X_t = Y_t$ , then*

$$\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{TV} \leq P[\tau > t]$$

In each chain, then, we begin with  $X_t$  started at a distribution of our choice, and  $Y_t$  started at stationarity. For any fixed (large)  $T$ , we will then couple  $X_t$  and  $Y_t$  so that they will have coupled by time  $T$  with high probability. Each coupling will have two phases: an initial phase from time 0 to time  $T_1$  in which  $X_t$  and  $Y_t$  get close with high probability, and a non-Markovian coupling phase from time  $T_1$  to time  $T = T_1 + T_2$  in which they are forced to collide. Unlike many coupling proofs, the time of interest  $T$  must be specified before constructing the coupling.

While the initial contraction phases are quite different for the two chains, the final coupling phase can be described in a unified way. The unifying device is the partition process  $P_t$  on set partitions of  $[n] = \{1, 2, \dots, n\}$ , introduced in [26] for a special case of the first sampler treated here (see that paper for details). This partition process contains some information about the coordinates  $\{i(t), j(t)\}_{t=T_1}^T$  used by  $Y_t$  throughout the entire process, and is the only source of information from the future that is used to construct the non-Markovian coupling. Critically, we don't use any information about the random variables  $\lambda(t)$  used at each step, which makes it trivial to check that the couplings constructed in this paper have the correct marginal distributions.

The process  $\{P_t\}_{t=T_1}^T$  will consist of a set of nested partitions of  $[n]$ ,  $P_{T_1} \leq P_{T_1+1} \leq \dots \leq P_T = \{\{1\}, \{2\}, \dots, \{n\}\}$ , where we say partition  $A$  is less than partition  $B$  if every element of partition  $B$  is a subset of an element of partition  $A$ . To construct  $P_t$ , we first look at the sequence of graphs  $G_t$  with vertex set  $[n]$  and edge set  $\{(i(s), j(s)) : s > t\}$ . Then let  $P_t$  consist of the connected components of  $G_t$ . While constructing  $P_t$ , we will also record a series of 'marked times'  $T = t_1 > t_2 > \dots > t_m$  and associated special subsets  $S(t_j, 1)$  and  $S(t_j, 2)$  of  $[n]$ . We will set  $t_1 = T$ , and then inductively set  $t_j = \sup\{t : t < t_{j-1}, P_{t-1} \neq P_t\}$ . Finally, note that if  $P_{t-1} \neq P_t$ , the only difference between them is that two elements of  $P_t$  have been merged into a single element in  $P_{t-1}$ . Label the set merged at time  $t_j$  with fewer elements  $S(t_j, 1)$ , and label the other one  $S(t_j, 2)$ . If both sets have the same number of elements, set  $S(t_j, 1)$  to be the one containing the smallest element (this is, of course, quite arbitrary).

We will be interested in the smallest time  $\tau$  such that  $P_{T-\tau} = [n]$ , a single block. From classical arguments (see e.g. chapter 7 of [3]),

**Lemma 4** (Connectedness). *For the Gibbs sampler on narrow matrices,*

$$P[\tau > (\frac{1}{2} + \epsilon)n \log(n)] \leq 2n^{-\epsilon}$$

The analogous lemma for the other example will be proved in Section 5, Lemma 7.

For both of our walks, we will use two types of coupling, the ‘proportional’ coupling and the ‘subset’ coupling. In both cases, we will set  $i(t)^x = i(t)^y$  and  $j(t)^x = j(t)^y$  at each step. In the proportional coupling, we will also set  $\lambda(t)^x = \lambda(t)^y$ .

To discuss the subset coupling, we must define the weight of  $X_t$  on a subset  $S \subset [n]$ , which we call  $w(X_t, S)$ . For the simplex walk, we define  $w(X_t, S) = \sum_{s \in S} X_t[s]$ . For narrow matrices, we define  $w(X_t, S) = \sum_{s \in S} X_t[s, 1]$ . The subset couplings associated with subset  $S \subset [n]$ , which are defined immediately prior to lemmas 18 and 8, will often set  $w(X_{t+1}, S) = w(Y_{t+1}, S)$ . We say that a subset coupling of subset  $S$  at time  $t$  succeeds if that equality holds; otherwise, we say it fails.

In each case, the coupling of  $X_t$  and  $Y_t$  during the non-Markovian coupling phase will be as follows. At marked times  $t_j$ , we will perform a subset coupling of  $X_{t_j}, Y_{t_j}$  with respect to  $S(t_j, 1)$ . At all other times, we will perform a proportional coupling. This leads to:

**Lemma 5** (Final Coupling). *Assume the non-Markovian coupling phase lasts from time  $T_1$  to  $T$ , that  $P_{T_1} = \{[n]\}$ , and that all subset couplings succeed. Then  $X_T = Y_T$ .*

*Proof.* Let  $\mathcal{F}_t$  denote the collection of equations  $w(X_t, S) = w(Y_t, S)$  for all  $S \in P_t$ . We will show by induction that the equations  $\mathcal{F}_t$  hold for all  $T_1 \leq t \leq T$ . At time  $T_1$ , we have  $w(X_{T_1}, [n]) = w(Y_{T_1}, [n]) = 1$ . By definition of the partition process, if  $t$  is not a marked time and all equations  $\mathcal{F}_t$  hold, then all equations  $\mathcal{F}_{t+1}$  also hold. In fact, this is true for any coupling of  $\lambda(t)^x, \lambda(t)^y$  at that step, not just the proportional coupling.

Assume  $t = t_j$  is a marked time, and that the equations  $\mathcal{F}_{t_j}$  hold. Then if  $\mathcal{F}_{t_j+1}$  don’t all hold, we must have that either  $w(X_{t_j+1}, S(t_j, 1)) \neq w(Y_{t_j+1}, S(t_j, 1))$  or  $w(X_{t_j+1}, S(t_j, 2)) \neq w(Y_{t_j+1}, S(t_j, 2))$ , since none of the terms in the other equations change. By assumption, all subset couplings have succeeded, so  $w(X_{t_j+1}, S(t_j, 1)) = w(Y_{t_j+1}, S(t_j, 1))$ . By construction,  $w(X_{t_j+1}, S(t_j, 2)) = w(X_{t_j+1}, S(t_j, 1) \cup S(t_j, 2)) - w(X_{t_j+1}, S(t_j, 1))$  and similarly for  $Y_{t_j+1}$ , so  $w(X_{t_j+1}, S(t_j, 2)) = w(Y_{t_j+1}, S(t_j, 2))$ . Thus, the inductive claim has been proved.

Finally, we note that if  $w(X_t, \{i\}) = w(Y_t, \{i\})$  for any singleton  $\{i\}$ , then  $X_t[i] = Y_t[i]$  for the sampler on the simplex (respectively  $X_t[i, j] = Y_t[i, j]$  for  $j \in \{1, 2\}$  for the other sampler). Since  $P_T = \{\{1\}, \{2\}, \dots, \{n\}\}$ , this proves the lemma. ■

So, in both cases, showing that all subset couplings succeed with high probability is sufficient to show that coupling has succeeded.

#### 4. CONTRACTION FOR GIBBS SAMPLERS ON THE SIMPLEX WITH GEOMETRY

In this section, we prove a contraction lemma for Gibbs samplers on the simplex associated with a group  $G$  and symmetric generating set  $R$  of  $G$  (that is,  $R^{-1} = R$ ), where  $|G| = n$ ,  $|R| = m$ , and  $id$  is the identity element of  $G$ . We recall briefly some definitions. We write  $\Delta_G = \{X \in \mathbb{R}^G | X[g] \geq 0, \sum_{g \in G} X[g] = 1\}$ . If  $X_t \in \Delta_G$  is a copy of the Markov chain, we take a step by choosing  $g \in G$ ,  $r \in R$  and  $\lambda \in [0, 1]$  uniformly and setting  $X_{t+1}[g] = \lambda(X_t[g] + X_t[gr])$ ,  $X_{t+1}[gr] = (1 - \lambda)X_t[g] + X_t[gr]$ , and for all other entries  $X_{t+1}[h] = X_t[h]$ . This walk is closely related to a slow simple random walk on the group. In particular, we let  $Z_t \in G$  be the random walk that evolves by choosing at each time step a group element  $g \in G$  and generator  $r \in R$  uniformly at random, and setting  $Z_{t+1} = Z_t r$  if  $Z_t = g$ , and  $Z_{t+1} = Z_t$  otherwise.

Let  $\hat{K}$  be the transition kernel associated with the random walk  $Z_t$ . Since  $R$  is symmetric, the random walk is reversible, so  $\hat{K}$  can be written in a basis of orthogonal eigenvectors with real eigenvalues  $1 = \hat{\lambda}_1 > \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_n \geq -1$ . Since it is  $\frac{1}{2}$ -lazy, all eigenvalues are in fact nonnegative. Let  $\hat{\gamma} = 1 - \hat{\lambda}_2$  be the spectral gap of  $\hat{K}$ . In this section we will show that

**Lemma 6** (Contraction Estimate for Gibbs Sampler on Cayley Graphs). *Let  $X_t, Y_t$  be two copies of the Gibbs sampler on the simplex associated with  $G$  and  $R$ , with joint distribution given by a proportional coupling at each step. Then*

$$E[||X_t - Y_t||_2^2] \leq 4ne^{-\lfloor \frac{t\hat{\gamma}}{8} \rfloor}$$

*Proof.* We will construct an auxilliary Markov chain on  $G$  associated with  $X_t$ , and compare it to the standard random walk  $Z_t$ . Let  $X_t, Y_t$  be two copies of the walk, and couple them at each step with the proportional coupling. For  $h \in G$ , let  $S_t^h = \sum_{g \in G} (X_t[g] - Y_t[g])(X_t[hg] - Y_t[hg])$ . We will analyze the evolution of the vector  $S_t = (S_t^{id}, \dots)$ .

There are three cases to analyze:  $h \in R$ ,  $h \notin R$  and  $h \neq id$ , and  $h = id$ . Let  $\mathcal{F}_t$  be the  $\sigma$ -algebra generated by  $X_s$  and  $Y_s$ ,  $0 \leq s \leq t$ . For case 1, we have

$$\begin{aligned}
E[S_{t+1}^h | \mathcal{F}_t] &= (1 - \frac{4}{n} + \frac{2}{mn})S_t^h \\
&+ \frac{1}{2mn} \sum_{i \in G} \sum_{r \in R, r \neq h, h^{-1}} [(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])(X_t[hi] - Y_t[hi]) \\
&+ (X_t[ri] + X_t[i] - Y_t[ri] - Y_t[i])(X_t[hri] - Y_t[hri]) \\
&+ (X_t[h^{-1}i] - Y_t[h^{-1}i])(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri]) \\
&+ (X_t[h^{-1}ri] - Y_t[h^{-1}ri])(X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])] \\
&+ \frac{2}{mn} \sum_{i \in G} [\frac{1}{6}(X_t[i] + X_t[hi] - Y_t[i] - Y_t[hi])^2 \\
&+ (X_t[i] - Y_t[i])(X_t[hi] - Y_t[hi]) + (X_t[i] - Y_t[i])(X_t[h^2i] - Y_t[h^2i])] \\
&= (1 - \frac{2}{n} + \frac{2}{3mn})S_t^h + \frac{2}{3mn}S_t^{id} + \frac{2}{mn}S_t^{h^2} \\
&+ \frac{1}{2mn} \sum_{r \in R, r \neq h, h^{-1}} (S_t^{hr^{-1}} + S_t^{hr} + S_t^{rh} + S_t^{rh^{-1}})
\end{aligned}$$

and we note that the sum of the coefficients is  $1 - \frac{2}{3mn}$ . For case 2, we have

$$\begin{aligned}
E[S_{t+1}^h | \mathcal{F}_t] &= (1 - \frac{4}{n})S_t^h + \frac{2m}{mn}S_t^h \\
&+ \frac{1}{2mn} \sum_{r \in R} (S_t^{hr^{-1}} + S_t^{hr} + S_t^{rh} + S_t^{rh^{-1}}) \\
&= (1 - \frac{2}{n})S_t^h + \frac{1}{2mn} \sum_{r \in R} (S_t^{hr^{-1}} + S_t^{hr} + S_t^{rh} + S_t^{rh^{-1}})
\end{aligned}$$

where the sum of the coefficients is 1. Finally, in case 3, we have

$$\begin{aligned}
E[S_{t+1}^{id} | \mathcal{F}_t] &= (1 - \frac{2}{n})S_t^{id} + \frac{2}{3mn} \sum_{r \in R} \sum_{i \in G} (X_t[i] + X_t[ri] - Y_t[i] - Y_t[ri])^2 \\
&= (1 - \frac{2}{3n})S_t^{id} + \frac{4}{3mn} \sum_{r \in R} S_t^r
\end{aligned}$$

and here the sum of the coefficients is  $1 + \frac{2}{3n}$ . If we rewrite  $U_t^{id} = \frac{1}{2}S_t^{id}$ , and otherwise  $U_t^g = S_t^g$ , then we find the following transformations. For case 1, we have

$$(2) \quad E[U_{t+1}^h | \mathcal{F}_t] = (1 - \frac{2}{n} + \frac{2}{3mn})U_t^h + \frac{4}{3mn}U_t^{id} + \frac{2}{mn}U_t^{h^2} \\ + \frac{1}{2mn} \sum_{r \in R, r \neq h, h^{-1}} (U_t^{hr^{-1}} + U_t^{hr} + U_t^{rh} + U_t^{rh^{-1}})$$

For case 2, we have

$$(3) \quad E[U_{t+1}^h | \mathcal{F}_t] = (1 - \frac{2}{n})U_t^h + \frac{1}{2mn} \sum_{r \in R} (U_t^{hr^{-1}} + U_t^{hr} + U_t^{rh} + U_t^{rh^{-1}})$$

Finally, in case 3, we have

$$(4) \quad E[U_{t+1}^{id} | \mathcal{F}_t] = (1 - \frac{2}{3n})U_t^{id} + \frac{2}{3mn} \sum_{r \in R} U_t^r$$

where the sum of the coefficients is now 1 in all three cases. In particular, the equations (2) to (4) now define a Markov chain on  $G$ . From equation (2), this random walk sends the identity to itself with probability  $1 - \frac{2}{3n}$ , and to a uniformly chosen element of  $S$  with the remaining probability; Equations (3) and (4) describe transitions for  $h \in R$  and  $h \notin R$  respectively. Call the transition kernel  $K$ .

Before analyzing the chain, we note that  $\sum_{i \in G} (X_t[i] - Y_t[i]) = 0$ , and so

$$0 = \left( \sum_{i \in G} (X_t[i] - Y_t[i]) \right)^2 \\ = \sum_{i \in G} (X_t[i] - Y_t[i])^2 + \sum_{i \neq j} (X_t[i] - Y_t[i])(X_t[j] - Y_t[j]) \\ = S_t^{id} + \sum_{h \neq id} S_t^h$$

From this calculation, if  $\langle v, (2, 1, 1, \dots, 1) \rangle = 0$ , then  $\langle Kv, (2, 1, 1, \dots, 1) \rangle = 0$  as well. By direct computation,  $\pi = \frac{1}{n+1}(2, 1, 1, \dots, 1)$  is a reversible measure for  $K$ . It is also clear that the distribution  $\hat{\pi} = \frac{1}{n}(1, 1, \dots, 1)$  is the reversible measure for  $\hat{K}$ .

We are now ready to compare the chains. Recall from [10] that the Dirichlet form associated to a Markov chain with transition kernel  $Q$  and stationary distribution  $\nu$  is given by

$$\mathcal{E}(\phi) = \frac{1}{2} \sum_{h, g \in G} \nu(g) Q(g, h) (\phi(x) - \phi(y))^2$$

Let  $\mathcal{E}$  and  $\hat{\mathcal{E}}$  be the Dirichlet forms associated with  $K$  and  $\hat{K}$  respectively. Then by comparing terms, it is clear that  $\mathcal{E}(\phi) \geq \frac{1}{4}\hat{\mathcal{E}}(\phi)$  for any  $\phi$  and  $\frac{\pi}{\hat{\pi}}, \frac{\hat{\pi}}{\pi} \leq 2$ . By e.g. Lemma 13.12 of [15], this implies  $\gamma \geq \frac{1}{8}\hat{\gamma}$ .



Recall that if  $\langle v, \pi \rangle = 0$ , then  $\langle K^m v, \pi \rangle = 0$  as well. In particular  $K$  applied to the subspace orthogonal to  $\pi$  has  $L^2 \rightarrow L^2$  operator norm at most  $1 - \gamma$ . Thus, we have for any  $v$  in that subspace

$$\|K^m v\|_2 \leq e^{-\lfloor \gamma m \rfloor} \|v\|_2$$

going back to our original situation, we are interested in the vector  $(S_t^g)$ . At time 0,  $S_0^{id} \leq 2$ , and by Cauchy-Schwarz  $|S_0^h| \leq 4$ . Thus,  $\|U_0^g\|_2 \leq 4n$ , and of course  $|S_t^{id}| \leq \|S_t^g\|_2 \leq \|U_t^g\|_2$ . So we find that

$$E[|S_t^{id}|] \leq 4ne^{-\lfloor \frac{t\hat{\gamma}}{8} \rfloor}$$

which is the contraction estimate in Lemma 6. ■

## 5. COUPLING FOR GIBBS SAMPLERS ON THE SIMPLEX WITH GEOMETRY

Having shown contraction, we must now show convergence in total variation distance. First, the analogue to Lemma 4:

**Lemma 7** (Connectedness for Gibbs Sampler on Cayley Graphs). *Let  $\tau$  be as defined immediately before Lemma 4 and let  $\hat{\gamma}$  be as defined immediately before Theorem 1. Then for  $t > 8 \frac{(C+3)\log(n)}{\hat{\gamma}}$ , we have*

$$P[\tau > t] \leq 2n^{-C}$$

*Proof.* We consider a graph-valued process  $G_t$ , where  $G_0$  is a graph with no edges, and vertex set equal to the group  $G$ . To construct  $G_{t+1}$  from  $G_t$ , choose elements  $g \in G$  and  $r \in R$  uniformly at random, and add the edge  $(g, gr)$  if it isn't already in  $G_t$ . We note that  $\tau > t$  if and only if  $G_t$  is not connected, so we would like to estimate the time at which  $G_t$  becomes connected.

First, fix two elements  $x, y \in G$ . We'd like to see if  $x, y$  are in the same component of  $G_t$ . To do so, let  $X_t, Y_t$  be two copies of the Gibbs sampler described in the last section, with  $X_0 = x, Y_0 = y$ . Couple  $X_t, Y_t$  and  $G_t$  by the proportional coupling. Then assume  $x, y$  are in different components  $C_x, C_y$  at time  $t$ . We would have

$$\begin{aligned} \sum_g |X_t[g] - Y_t[g]|^2 &\geq \sum_{g \in C_x} \frac{1}{|C_x|^2} + \sum_{g \in C_y} \frac{1}{|C_y|^2} \\ &\geq \frac{4}{n} \end{aligned}$$

By Markov's inequality, then,

$$P[C_x \neq C_y] \leq \frac{n}{4} E\left[\sum_g |X_t - Y_t|^2\right]$$

and so, by standard union bound for fixed  $x$  over all  $y$ , if  $A_t$  is the event that  $G_t$  is disconnected,

$$\begin{aligned} P[A_t] &\leq \frac{n^2}{4} \sup_{\mu, \nu} E\left[\sum_g |X_t - Y_t|^2 \mid X_0 = \mu, Y_0 = \nu\right] \\ &\leq 2n^3 e^{-\lfloor \frac{t\hat{\gamma}}{8} \rfloor} \end{aligned}$$

where the last inequality is due to Lemma 6. ■

Next, we define subset couplings and discuss success probabilities for this walk. Fix points  $X_t, Y_t$ , subset  $S \subset [n]$  and updated coordinates  $i = i(t) \in S, j = j(t) \notin S$ . The next step is to construct a pair of uniform random variables  $\lambda_x = \lambda(t)^x$  and  $\lambda_y = \lambda(t)^y$  with which to update the chains  $X_t$  and  $Y_t$  respectively. Assume first that  $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} < 1$ , and choose  $\lambda_y$  uniformly in  $[0, 1]$ . Then set

$$(5) \quad \lambda_x = \lambda_y \frac{Y_t[i] + Y_t[j]}{X_t[i] + X_t[j]} + \frac{1}{X_t[i] + X_t[j]} \sum_{s \in S/\{i\}} (Y_t[s] - X_t[s])$$

if that results in a value between 0 and 1. Otherwise, choose  $\lambda_x$  independently of  $\lambda_y$ , according to the density:

$$(6) \quad f(\lambda) = C \left( 1 - \frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} \mathbf{1}_{g(\lambda) \in [0, 1]}(\lambda) \right)$$

where  $C^{-1} = \int_0^1 f(\lambda) d\lambda$  is a normalizing constant, and

$$g(\lambda) = \lambda \frac{Y_t[i] + Y_t[j]}{X_t[i] + X_t[j]} + \frac{1}{X_t[i] + X_t[j]} \sum_{s \in S/\{i\}} (Y_t[s] - X_t[s])$$

From the assumption that  $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} < 1$ , it is easy to see that  $f$  really is a density on  $[0, 1]$ . From its construction as a remainder density, it is easy to check that under this coupling,  $\lambda_x$  is uniformly distributed on  $[0, 1]$ . If  $\frac{X_t[i] + X_t[j]}{Y_t[i] + Y_t[j]} > 1$ , an analogous construction will work. More precisely, in this case choose  $\lambda_x$  first, and then choose  $\lambda_y$  to satisfy equation 5 if the result is in  $[0, 1]$ , rather than choosing  $\lambda_y$  first. If the result is not in  $[0, 1]$ , then choose  $\lambda_y$  according to its remainder measure, given by equation (6) with  $X_t$  and  $Y_t$  flipped and  $g$  replaced by  $g^{-1}$ . Note that if equation 5 is satisfied, then  $w(S, X_{t+1}) = w(S, Y_{t+1})$ .

For a pair of points  $(x, y)$  in the simplex, a pair of update entries  $(i, j)$ , and a subset  $S \subset [n]$  of interest such that  $i \in S$  and  $j$  not in  $S$ , we define  $p(x, y, i, j, S)$  to be the probability that the associated subset coupling succeeds. Then the following lemma from [26] gives a lower bound on this probability:

**Lemma 8** (Subset Coupling). *For a pair of vectors  $(x, y)$  satisfying  $\sup_i |x_i - y_i| \leq n^{-e}$  and  $\inf_i x_i, \inf_i y_i \geq n^{-b}$ , for  $e > b$ , we have for all sufficiently large  $n$  that  $p(x, y, i, j, S) \geq 1 - 2n^{b+1-e}$  uniformly in  $S$  and possible  $i, j$ .*

In general, it is possible to choose  $x, y, i, j, S$  so that the probability of success is 0 under any coupling, and the lemma is quite restrictive. Having bounded the probability of failure when  $X_t, Y_t$  are close, we must show that they remain close with high probability. Define for  $v \in \mathbb{R}^n$  and  $S \subset [n]$  the quantity  $\|v\|_{1,S} = \sum_{s \in S} |v[s]|$ . Then:

**Lemma 9** (Smallness). *Let  $X_t, Y_t$  be coupled as described in Section 3, and assume that  $P_{T_1} = \{[n]\}$ , that all subset couplings up to time  $t$  have succeeded, and that  $\|X_{T_1} - Y_{T_1}\|_1 < \epsilon$ . Then  $\|X_t - Y_t\|_{1,S} < \epsilon$  for every  $S$  in  $P_t$ .*

*Proof.* There are two types of coupling to take care of. For a proportional coupling between  $i$  and  $j$ , we note that the error  $\Delta$  satisfies:

$$\begin{aligned} \Delta &= |X_{t+1}[i] - Y_{t+1}[i]| + |X_{t+1}[j] - Y_{t+1}[j]| \\ &= \lambda(t)|X_t[i] + X_t[j] - Y_t[i] - Y_t[j]| + (1 - \lambda(t))|X_t[i] + X_t[j] - Y_t[i] - Y_t[j]| \\ &\leq |X_t[i] - Y_t[i]| + |X_t[j] - Y_t[j]| \end{aligned}$$

Since  $i$  and  $j$  always connect elements of the same set in  $P_t$ , this shows that proportional couplings never increase  $\|X_t - Y_t\|_{1,S}$ .

Otherwise, assume that at time  $t$  we had a successful subset coupling between subsets  $S(1), S(2)$  along edge  $i, j$ , with  $i$  in  $S(1)$  and  $j$  in  $S(2)$ . Then we note that

$$\begin{aligned} X_{t+1}[i] - Y_{t+1}[i] &= \sum_{s \in S(1) \setminus i} (Y_t[s] - X_t[s]) \\ &= Y_t[i] - X_t[i] + \sum_{s \in S(2)} (X_t[s] - Y_t[s]) \end{aligned}$$

and so

$$|X_{t+1}[i] - Y_{t+1}[i]| \leq |X_t[i] - Y_t[i]| + \|X_t - Y_t\|_{1,S(2)}$$

which immediately implies that

$$\|X_{t+1} - Y_{t+1}\|_{1,S(2)} \leq \|X_t - Y_t\|_{1,S(1) \cup S(2)}$$

Inductively, this shows that  $\|X_{t+1} - Y_{t+1}\|_{1,S} \leq \|X_0 - Y_0\|_1$ . ■

Related to this, the following lemma from chapter 13 of [1] shows that  $X_t, Y_t$  rarely have entries close to 0:

**Lemma 10** (Largeness).  $P[\inf_{1 \leq i \leq n} \inf_{0 \leq t \leq T} Y_t[i] \leq n^{-b}] \leq Tn^{3-b}$

This lets us complete the calculation. Assume that the initial contractive phase is of length  $T_1 = \frac{8C_1}{\hat{\gamma}} \log(n)$ , and that the second coupling phase is of length  $T_2 = \frac{8C_2}{\hat{\gamma}} \log(n)$ .

By Lemma 6,  $E[\sum_{g \in G} (X_{T_1}[g] - Y_{T_1}[g])^2] \leq 4n^{1-C_1}$ , and combining this with Markov's inequality and the bound  $\|V\|_1 \leq \sqrt{2n}\|V\|_2$ , we have  $P[\sum_{g \in G} |X_{T_1}[g] - Y_{T_1}[g]| \geq n^{-a}] \leq n^{1+a-\frac{1}{2}C_1}$ .

By Lemma 10,  $P[\inf_{0 \leq t \leq \frac{n}{\hat{\gamma}}, g \in G} Y_t[g] \leq n^{-b}] \leq \frac{1}{\hat{\gamma}} n^{3-b}$ . By Lemma 7, the probability that  $P_{T_1}$  consists of a single block is at least  $1 - 2n^{3-C_2}$ . Finally, by lemmas 8 and 9 the probability that any of the subset couplings fail while  $\inf_{0 \leq t \leq \frac{n}{\hat{\gamma}}, g \in G} Y_t[g] \geq n^{-b}$  and  $\sum_{g \in G} |X_{T_1}[g] - Y_{T_1}[g]| \geq n^{-a}$  is less than  $2n^{b+2-a}$ . By Lemma 5,  $X_T = Y_T$  unless one of the subset couplings fails or  $\tau > T_2$ . As written, the sum of the probabilities that one of these two events don't occur is thus at most  $8n^{1+a-\frac{1}{2}C_1} + \frac{1}{\hat{\gamma}} n^{3-b} + 2n^{3-C_2} + 2n^{b+2-a}$ . It is easy to show that  $\hat{\gamma} \geq \frac{1}{2n^4}$  for simple random walk on any Cayley graph (e.g. by naive bounds with Theorem 13.14 of [15]), which changes the second term to  $2n^{7-b}$ . To come close to minimizing this, if  $T = 8(7x+23)n \log(n)$ , set  $b = x+7$ ,  $a = 2x+9$ ,  $C_1 = 6x+20$ , and  $X_2 = x+3$ . For this, find that the probability of failure is at most  $14n^{-x}$ .

Thus, we have shown that for the simplex walk, for  $t > 8(7x+23)\frac{\log(n)}{\hat{\gamma}} - 8\frac{\log(\hat{\gamma})}{\hat{\gamma}}$ ,

$$(7) \quad \|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{TV} \leq 14n^{-x}$$

which proves the upper bound in Theorem 1.

## 6. LOWER BOUNDS FOR GIBBS SAMPLERS ON THE SIMPLEX WITH GEOMETRY

In this section, we prove lower bounds on the mixing time of the Gibbs sampler on the simplex. The results are similar to those of [22], though the method is different and elementary. Begin by calculating

$$\begin{aligned} E[X_{t+1}[g]|X_t] &= (1 - \frac{2}{n})X_t[g] + \frac{2}{n} \frac{1}{m} \sum_{r \in R} \frac{1}{2} (X_t[g] + X_t[gr]) \\ &= (1 - \frac{1}{n})X_t[g] + \frac{1}{n} \frac{1}{m} \sum_{r \in R} X_t[gr] \end{aligned}$$

In particular, let  $K$  be the transition matrix on  $G$  given by  $K[g, g] = 1 - \frac{1}{n}$ , and  $K[g, gr] = \frac{1}{nm}$  for  $r \in R$ . This is the standard 'edge'-based random walk on  $G$  with generating set  $R$  described above. This calculation shows that  $E[X_t] = K^t X_0$ . Note that this is not the same  $K$  as was used earlier in the section while proving the upper bound on the mixing time. By the earlier assumptions on  $R$ ,  $K$  is reversible with respect to the uniform measure on  $G$ . Furthermore, it is orthogonally diagonalizable with real eigenvalues  $1 = \beta_1 > \beta_2 \geq \dots \geq \beta_n \geq 0$ .

Next, let  $v$  be an eigenvector of  $K$  with eigenvalue  $\beta_2$ , normalized so that  $\|v\|_2 = 1$  and  $\|v - Kv\|_2 = \gamma$ , the spectral gap of  $K$ . Let  $\Pi$  be the collection of vectors with nonnegative entries summing to 1, and let  $w \in \Pi$  maximize the inner product  $\langle v, w \rangle$  among such vectors; such a vector exists by the compactness of  $\Pi$ . Let  $X_t$  be a copy of the Markov chain begun from  $X_0 = w$ , then  $E[\langle X_t, v \rangle] = (1 - \gamma)^t \langle w, v \rangle$ . On the other hand, if  $A_{t,d}$  is the event that  $\langle X_t, v \rangle > d$ ,

$$\begin{aligned} E[\langle X_t, v \rangle] &= E[\langle X_t, v \rangle 1_{A_{t,d}}] + E[\langle X_t, v \rangle 1_{A_{t,d}^c}] \\ &\leq \langle X_0, v \rangle P[A_{t,d}] + d \end{aligned}$$

where the second inequality takes advantage of the maximality of  $X_0$ . Thus,

$$P[A_{t,d}] \geq \frac{E[\langle X_t, v \rangle] - d}{\langle X_0, v \rangle}$$

putting the two inequalities together,

$$P[A_{t,d}] \geq (1 - \gamma)^t - \frac{d}{\langle X_0, v \rangle}$$

The next step is to prove that  $\langle X_0, v \rangle \geq \frac{2}{\sqrt{n}}$ . Let  $P \subset [n]$  be the collection of indices so that  $v[p] \geq 0$  for  $p \in P$ . Without loss of generality, assume  $\sum_{p \in P} v_p^2 > \frac{1}{2}$ . Now set  $\lambda^{-1} = \sum_{p \in P} v_p \leq \sqrt{n}$ . Then consider the distribution given by  $\mu_p = \lambda v_p$  for  $p \in P$ , and  $\mu_p = 0$  for  $p \notin P$ :

$$\begin{aligned} \langle \mu, v \rangle &= \lambda \sum_{p \in P} v_p^2 \\ &\geq \frac{2}{\sqrt{n}} \end{aligned}$$

and so

$$P[A_{t,d}] \geq (1 - \gamma)^t - \frac{1}{2} d \sqrt{n}$$

Now, let  $Y \in \Delta_G$  be chosen according to the uniform distribution. Then  $E[\langle Y, v \rangle] = 0$ , and

$$\begin{aligned} E[\langle Y, v \rangle^2] &= E\left[\left(\sum_{i \in G} Y[i] v_i\right)^2\right] \\ &= \sum_{i \in G} v_i^2 E[Y[i]^2] + \sum_{i \neq j \in G} v_i v_j E[Y[i] Y[j]] \\ &\leq \frac{2}{n^2} + 0 \end{aligned}$$

So, by Chebyshev's inequality,  $P[\langle Y, v \rangle > d] \leq \frac{2}{d^2 n^2}$ . Putting this together with the inequality above, letting  $d = n^{-\frac{5}{6}}$  and defining  $P[A_{\infty, d}] = \lim_{t \rightarrow \infty} P[A_{t, d}]$ ,

$$P[A_{t, d}] - P[A_{\infty, d}] \geq \gamma^t - 3n^{-\frac{1}{3}}$$

And the lower bound follows immediately.

## 7. CONTRACTION AND NARROW MATRICES

We begin with some quick observations about the geometry of our space. It is the part of an  $(n - 1)$ -dimensional affine subspace of  $\mathbb{R}^{2n}$  that lies in the upper orthant. Our updates are in fact moves along 1-dimensional pieces of this subspace, even though we are updating four entries. While the original motivation for this sampler comes from statistics (see e.g. [6]), it is being treated here primarily as an example of a chain that is somewhere between the standard Gibbs sampler on the simplex and an analogous Gibbs sampler on doubly-stochastic matrices or Kac's famous walk on the orthogonal group. The former was analyzed by the author in [26], using a simpler non-Markovian coupling argument. Matching bounds on Total variation mixing time are not known for either the Gibbs sampler on doubly-stochastic matrices or Kac's walk. The best such bounds to date can be found in [27] and [14] respectively. Both mixing bounds are polynomials with small but probably incorrect degrees, and both are based on much more complicated non-Markovian coupling arguments.

In this section, we will prove contractivity estimates for the Gibbs sampler on narrow matrices. The work will be done in a combination of the  $L^2$  metric,  $\|X_t - Y_t\|_2^2 = \sum_{i=1}^n (X_t[i, 1] - Y_t[i, 1])^2$ , and the  $L^1$  metric,  $\|X_t - Y_t\|_1 = \sum_{i=1}^n |X_t[i, 1] - Y_t[i, 1]|$ . The following main estimate will be proved by a sequence of lemmas:

**Lemma 11** (Weak Convergence on Narrow Matrices). *If  $X_t$  and  $Y_t$  are coupled under the proportional coupling until time  $T_1 = (10k + 10.5)n \log(n)$ , then*

$$P[\|X_{T_1} - Y_{T_1}\|_1 \geq \epsilon] \leq 3\epsilon^{-1} n^{-k}$$

We begin with a non-rigorous description of the proof strategy for this lemma, which takes the rest of this section. The lemma is a contraction result, and it will be proved using a variant of the path-coupling argument introduced in [4]. In path-coupling arguments, the goal is to couple  $X_t$  and  $Y_t$  by constructing an interpolating chain,  $X_t = Z_t^{(0)}, Z_t^{(1)}, \dots, Z_t^{(m)} = Y_t$  so that  $d(X_0, Y_0) \sim \sum_{j=1}^m d(Z_0^{(j-1)}, Z_0^{(j)})$  for some metric  $d$ . We would then show that, in general,  $E[d(Z_t^{(j-1)}, Z_t^{(j)})] \leq \alpha^t d(Z_0^{(j-1)}, Z_0^{(j)})$  for some  $0 < \alpha < 1$ . In most coupling arguments, we find such an  $\alpha$  that holds for all pairs  $Z_t^{(j)}, Z_t^{(j+1)}$  associated with a typical pair  $X_t$  and  $Y_t$ ; this immediately gives an estimate of  $E[d(X_t, Y_t)] \leq \alpha^t \sum_{j=1}^m d(Z_0^{(j-1)}, Z_0^{(j)}) \sim \alpha^t d(X_0, Y_0)$  for most starting pairs  $X_0, Y_0$ . This is converted into a bound on all starting pairs by adding a small period, known as the 'burn-in', to remove any bad features of  $X_0$  and  $Y_0$  with high

probability. We will also need a burn-in period, and divide  $T$  into an initial burn-in of length  $T_1$  and a contractive period of length  $2T_2$ .

In our argument, we show a contraction estimate only for most pairs  $Z_t^{(j-1)}, Z_t^{(j)}$  in an interpolation between typical chains  $X_t, Y_t$ . While this generally causes arguments for chains over finite spaces to fail completely, it leads to only slightly worse bounds for this and (we conjecture) other chains on continuous spaces. To be more precise, Lemma 12 gives an  $L^2$  contraction coefficient of  $(1 - \frac{2}{3n})$  for sufficiently nearby points. On the other hand, inequality (13) indicates that under the proportional coupling,  $E[d(X_{2t}, Y_{2t})] \leq C(n)(1 - \frac{2}{3n})^t$  for some constant  $C(n)$ . In particular, the global contraction estimate and the local contraction estimate asymptotically differ by a factor of only 2.

Having discussed the big picture, we will now begin the proof by making some basic remarks about the chain, beginning with an alternative description of the transition probabilities. Define  $\delta_t[i, j] = 2 - X_t[i, 1] - X_t[j, 1]$ , and  $\epsilon_t[i, j] = 2 - Y_t[i, 1] - Y_t[j, 1]$ . Then a step of the chain can be defined in the following way. Choose  $i, j$  as before, and choose  $\lambda \stackrel{D}{=} U[0, 1]$ . If  $\delta_t[i, j] \geq 0$ , then we update according to:

$$(8) \quad \begin{aligned} X_{t+1}[i, 1] &= \lambda(2 - \delta_t[i, j]) \\ X_{t+1}[j, 1] &= (1 - \lambda)(2 - \delta_t[i, j]) \\ X_{t+1}[i, 2] &= 2(1 - \lambda) + \lambda\delta_t[i, j] \\ X_{t+1}[j, 2] &= 2\lambda + (1 - \lambda)\delta_t[i, j] \end{aligned}$$

If  $\delta_t[i, j] < 0$ , we update according to:

$$(9) \quad \begin{aligned} X_{t+1}[i, 1] &= 2\lambda - (1 - \lambda)\delta_t[i, j] \\ X_{t+1}[j, 1] &= 2(1 - \lambda) - \lambda\delta_t[i, j] \\ X_{t+1}[i, 2] &= (1 - \lambda)(2 + \delta_t[i, j]) \\ X_{t+1}[j, 2] &= \lambda(2 + \delta_t[i, j]) \end{aligned}$$

Note that in both cases, a larger value of  $\lambda$  means a larger value of  $X_{t+1}[i, 1]$ . We are now ready to describe the proportional coupling: as in the simplex case, we choose the same value of  $\lambda$  for both chains in the above representation.

We will need two initial contractivity lemmas, dealing with  $L^2$  contractivity for nearby points and a general but poor bound in  $L^1$ . The first is:

**Lemma 12** ( $L^2$  Contractivity). *If  $X_t, Y_t$  are coupled under the proportional coupling for time  $0 \leq t \leq T_1$ , and  $\mathbf{1}_A$  is the indicator function  $\delta_t[i, j]\epsilon_t[i, j] \geq 0$  for all  $1 \leq i, j \leq n$  and all  $0 \leq t \leq T_1$ , then*

$$E[\|X_{T_1} - Y_{T_1}\|_2^2 \mathbf{1}_A] \leq (1 - \frac{2}{3n})^{T_1} \|X_0 - Y_0\|_2^2$$

*Proof.* We begin the proof by calculating the change in the  $L^2$  norm during a single move. Let  $F_t[i, j]$  be the event that coordinates  $i, j$  are updated at time  $t$ . We find:

$$\begin{aligned}
\Delta_t &\equiv E[(X_{t+1}[i, 1] - Y_{t+1}[i, 1])^2 + (X_{t+1}[j, 1] - Y_{t+1}[j, 1])^2 \\
&\quad + (X_{t+1}[i, 2] - Y_{t+1}[i, 2])^2 + (X_{t+1}[j, 2] - Y_{t+1}[j, 2])^2 | F_t[i, j]] \\
&= 2E[(\lambda(2 - \delta_t[i, j]) - \lambda(2 - \epsilon_t[i, j]))^2 + (\lambda\delta_t[i, j] - \lambda\epsilon_t[i, j])^2 | F_t[i, j]] \\
&= \frac{4}{3}(\delta_t[i, j] - \epsilon_t[i, j])^2
\end{aligned}$$

It would be nice to calculate the sums of terms like  $(\epsilon_t[i, j] - \delta_t[i, j])^2$  in terms of sums of terms like  $(X_t[i, j] - Y_t[i, j])^2$ . Fortunately, as in the simplex case, these are easy to relate. We first note that

$$\begin{aligned}
0 &= \left( \sum_{i=1}^n X_t[i, 1] - Y_t[i, 1] \right)^2 \\
&= \sum_{i=1}^n (X_t[i, 1] - Y_t[i, 1])^2 + 2 \sum_{i < j} (X_t[i, 1] - Y_t[i, 1])(X_t[j, 1] - Y_t[j, 1])
\end{aligned}$$

If  $\delta_t[i, j]\epsilon_t[i, j] \geq 0$  for all  $i, j$ , we can write the first line of the following computation:

$$\begin{aligned}
\sum_{i \neq j} (\delta_t[i, j] - \epsilon_t[i, j])^2 &= \sum_{i \neq j} (X_t[i, 1] + X_t[j, 1] - Y_t[i, 1] - Y_t[j, 1])^2 \\
&= \sum_{i \neq j} [(X_t[i, 1] - Y_t[i, 1])^2 + (X_t[j, 1] - Y_t[j, 1])^2 \\
&\quad + 2(X_t[i, 1] - Y_t[i, 1])(X_t[j, 1] - Y_t[j, 1])] \\
(10) \qquad &= (n-2) \sum_{j=1}^2 \sum_{i=1}^n (X_t[i, j] - Y_t[i, j])^2
\end{aligned}$$



Then the final contraction is given by:

$$\begin{aligned}
\Delta_t &= E[||X_{t+1} - Y_{t+1}||_2^2 | X_t, Y_t] \\
&= \frac{1}{n(n-1)} \sum_{k=1}^2 \sum_{m=1}^n \sum_{i \neq j} E[(X_{t+1}[m, k] - Y_{t+1}[m, k])^2 | F_t[i, j]] \\
&= \frac{1}{n(n-1)} \sum_{k=1}^2 \sum_{m=1}^n \left( \sum_{i, j \neq m} (X_t[m, k] - Y_t[m, k])^2 \right. \\
&\quad \left. + 2 \sum_{j \neq m} E[(X_{t+1}[m, k] - Y_{t+1}[m, k])^2 | F_t[i, j]] \right) \\
&= (1 - \frac{2}{n}) ||X_t - Y_t||_2^2 + \frac{1}{n(n-1)} \sum_{i \neq j} \frac{4}{3} (\delta_t[i, j] - \epsilon_t[i, j])^2 \\
&\leq (1 - \frac{2}{3n}) ||X_t - Y_t||_2^2
\end{aligned}$$

where the last line is due to equation (7). On the set  $A$ , iterating this inequality over  $t$  immediately implies the estimate in the statement of the lemma. On  $A^c$ , the expectation is of course exactly 0.  $\blacksquare$

**Lemma 13** ( $L^1$  Contractivity). *If  $X_t, Y_t$  are coupled under the proportional coupling for time  $0 \leq t \leq T_1$ , then*

$$(11) \quad ||X_{T_1} - Y_{T_1}||_1 \leq ||X_0 - Y_0||_1$$

*Proof.* Considering the cases  $\delta_t[i, j]\epsilon_t[i, j] \geq 0$  and  $\delta_t[i, j]\epsilon_t[i, j] \leq 0$ , this follows immediately by induction on  $t$  from applying the triangle inequality to the formulae (8) and (9).  $\blacksquare$

Having demonstrated contractivity for ‘nice’ pairs  $X_t$  and  $Y_t$ , we must now look at ‘typical’ pairs  $X_t$  and  $Y_t$ . The following burn-in lemma shows that, after a moderate number of steps,  $X_t$  and  $Y_t$  are unlikely to be too close to the boundary of our convex set.

**Lemma 14** (Burn-in). *For any starting position  $X_0$ ,*

$$P[\inf_{i,j} X_t[i, j] \leq n^{-k}], P[\sup_{i,j} X_t[i, j] \geq 2 - n^{-k}] \leq n^{-\frac{t}{7n \log(n)} + 6.5} + n^{-k+3}$$

*Proof.* Our proof will be via comparison to a Gibbs sampler on the simplex, studied by the author in [26]. Let  $X_t$  be a copy of our Gibbs sampler on 2 by  $n$  matrices, and let  $S_t$  be a Gibbs sampler on the simplex  $\Delta_n = \{S \in \mathbb{R}^n | S[i] \geq 0, \sum_{i=1}^n S[i] = 1\}$ . To make a move in this Gibbs sampler, choose distinct coordinates  $1 \leq i, j \leq n$  and  $0 \leq \lambda \leq 1$  uniformly at random, and update entry  $S_t[i]$  to  $\lambda(S_t[i] + S_t[j])$  and entry  $S_t[j]$  to  $(1 - \lambda)(S_t[i] + S_t[j])$ , keeping all other entries fixed. This is identical to the other sampler in this note, with generating set  $R = G \setminus \{id\}$ .

Since  $\sum_i S_0[i] = 1$ , for any given  $X_0$  it is possible to choose a corresponding  $S_0$  such that  $X_0[i, 1] \geq S_0[i]$  for all  $i$ , without the row sum condition interfering. Next, under our descriptions there is a natural proportional coupling of  $X_t$  and  $S_t$ , given by always choosing  $i, j$  and  $\lambda$  to be the same. We claim that under this coupling,  $X_t[i, 1] \geq S_t[i]$  for all  $t > 0$  and all  $1 \leq i \leq n$ . Assume inductively that this holds until time  $t$ , and that coordinates  $i, j$  are updated at time  $t$ . Using the representation in (8) and (9)

$$\begin{aligned} X_{t+1}[i, 1] &\geq \lambda \min(X_t[i, 1] + X_t[j, 1], 2) \\ &\geq \lambda \min(S_t[i] + S_t[j], 2) \\ &= \lambda(S_t[i] + S_t[j]) \\ &= S_{t+1}[i] \end{aligned}$$

Let  $S$  be drawn from the uniform distribution on the simplex. Then the above monotonicity tells us that

$$P[\inf_i X_t[i, 1] \leq n^{-k}] \leq \|\mathcal{L}(S_t) - U\|_{TV} + P[\inf_i S[i] \leq n^{-k}]$$

From [26],  $\|\mathcal{L}(S_t) - U\|_{TV} \leq n^{-\frac{t}{7n \log(n)} + 6.5}$  and  $P[\inf_i S[i] \leq n^{-k}] \leq n^{-k+3}$ . This gives a good bound on  $P[\inf_i X_t[i, 1] \leq n^{-k}]$ . Since all rows sum to 2, this gives the same bound on  $P[\sup_i X_t[i, 2] \geq 2 - n^{-k}]$ . Since there is symmetry between the top and bottom rows, this completes the proof.  $\blacksquare$

To analyze the second part of the coupling used to prove Lemma 11, we create an interpolating sequence between  $X_{T_1}$  and  $Y_{T_1}$ , for some  $T_1 = c_0 n \log(n)$  large enough that the burn-in lemma has taken effect. Since our sample space is convex, this will be simple. We define  $X_{T_1} = Z_{T_1}^1, Z_{T_1}^2, \dots, Z_{T_1}^l = Y_{T_1}$  so that  $\|Z_{T_1}^m - Z_{T_1}^{m+1}\|_1$  is very small, and so that all of the  $Z_{T_1}^l$  are in order along the line between  $X_{T_1}$  and  $Y_{T_1}$ . That is,  $Z_{T_1}^i$  is closer to  $X_{T_1}$  than  $Z_{T_1}^j$  if  $i < j$ . For the following lemma and later use in this paper, we define  $\delta_t^m[i, j] = 2 - Z_t^m[i, 1] - Z_t^m[j, 1]$ , analogously to the definition of  $\delta_t[i, j]$  in the preliminary calculations. We also define  $D[Z_t^{m_1}, Z_t^{m_2}] = |\{(i, j) : \delta_t^{m_1}[i, j] \delta_t^{m_2}[i, j] < 0\}|$ . Then:

**Lemma 15** (Interpolating Sequence). *The interpolating sequence described above satisfies:*

- (1)  $|\{m : D[Z_{T_1}^m, Z_{T_1}^{m+1}] \geq 1\}| \leq n^2$
- (2)  $\min(X_{T_1}[i, j], Y_{T_1}[i, j]) \leq Z_{T_1}^m[i, j] \leq \max(X_{T_1}[i, j], Y_{T_1}[i, j])$  for all  $i, j, m$ .

*Proof.* Part (2) follows from the fact that  $Z_{T_1}^m$  is on a line between  $X_{T_1}$  and  $Y_{T_1}$ , and hence all coordinates are between those of  $X_{T_1}$  and  $Y_{T_1}$ . For part (1), we observe that  $\delta_{T_1}^m[i, j]$  changes sign at most once as  $m$  changes, for any fixed pair  $i, j$ . The inequality follows since the number of pairs  $i, j$  is less than  $\frac{n^2}{2}$ .  $\blacksquare$

We are finally ready to show that that adjacent pairs in the interpolating sequence get closer, when the entire sequence is run under the proportional coupling. Let  $A[m, t]$  be the event that  $D[Z_s^m, Z_t^m] = 0$  for  $s \leq t$ . Then by lemmas 7 and 8, we have

$$\begin{aligned} E[\|Z_t^m - Z_t^{m+1}\|_2^2 1_{A[m, t]}] &\leq (1 - \frac{2}{3n})^t \|Z_0^m - Z_0^{m+1}\|_2^2 \\ \|Z_t^m - Z_t^{m+1}\|_1 &\leq \|Z_0^m - Z_0^{m+1}\|_1 \end{aligned}$$

which, combined with the Cauchy-Schwarz bound  $\|V\|_1 \leq \sqrt{2n}\|V\|_2$ , tells us that for  $n \geq 2$ ,

$$(12) \quad E[\|Z_{2t}^m - Z_{2t}^{m+1}\|_1] \leq n(1 - \frac{2}{3n})^t \|Z_0^m - Z_0^{m+1}\|_2 + P[A[m, 2t]^c] \|Z_0^m - Z_0^{m+1}\|_1$$

So, it remains to bound  $P[A[m, t]^c]$ . To do so, let  $G[t, m, k]$  be the event that  $Z_t^m[i, j]$  and  $Z_t^{m+1}[i, j]$  are all at least  $n^{-k}$  away from 0 and 2, and that  $D[Z_t^m, Z_t^{m+1}] = 0$ . Then we find that

**Lemma 16** (Chamber Occupancy).  $P[D[Z_{t+1}^m, Z_{t+1}^{m+1}] \geq 1 | G[t, m, k]] \leq n^{2+k} \|Z_0^m - Z_0^{m+1}\|_1$ .

*Proof.* To prove this, just note that at the next move,  $P[\delta_{t+1}^m[i, j] \delta_{t+1}^{m+1}[i, j] < 0]$  is bounded above by the ratio of the  $L^1$  distance between  $Z_{t+1}^m$  and  $Z_{t+1}^{m+1}$  to the total range that can be travelled. But the former is bounded above by  $\|Z_0^m - Z_0^{m+1}\|_1$ , and the latter is bounded below by  $n^{-k}$ . This, combined with a union bound over all  $\frac{n(n-1)}{2}$  pairs of distinct  $i, j$  gives the bound.  $\blacksquare$

Thus, using a union bound for Lemma 16 over time from 0 to  $t < n^2$ , as well as Lemma 14, we find that for any  $k > 2$ ,  $P[A[m, t]^c] \leq 2\|Z_0^m - Z_0^{m+1}\|_1 n^{4+k} + n^{4-k}$  after a burn in period of at least  $T_1 = (7k + 6.5)n \log(n)$ . Putting this together with inequality (12), we find that after the burn-in period,

$$\begin{aligned} E[\|Z_{T_1+2t}^m - Z_{T_1+2t}^{m+1}\|_1] &\leq n(1 - \frac{2}{3n})^t \|Z_{T_1}^m - Z_{T_1}^{m+1}\|_2 \\ &\quad + 2\|Z_{T_1}^m - Z_{T_1}^{m+1}\|_1 (n^{4+k} \|Z_{T_1}^m - Z_{T_1}^{m+1}\|_1 + n^{4-k}) \end{aligned}$$

By the triangle inequality,  $\|X_{2t} - Y_{2t}\|_1 \leq \sum_{m=1}^{l-1} \|Z_{2t}^m - Z_{2t}^{m+1}\|_1$ , so

$$\begin{aligned} E[\|X_{T_1+2t} - Y_{T_1+2t}\|_1] &\leq n(1 - \frac{2}{3n})^t \sum_{m=1}^{l-1} \|Z_{T_1}^m - Z_{T_1}^{m+1}\|_2 \\ &\quad + 2 \sum_{m=1}^{l-1} \|Z_{T_1}^m - Z_{T_1}^{m+1}\|_1 (n^{4+k} \|Z_{T_1}^m - Z_{T_1}^{m+1}\|_1 + n^{4-k}) \end{aligned}$$

We note that this inequality holds for any choice of  $k, l$  possibly depending on  $t$ . In particular, if  $T_2 = \frac{3}{2}(A+1)n \log(n)$ , then choosing  $k = A+3$  and  $l = n^{A+7}$  gives

$$(13) \quad E[\|X_{T_1+2T_2} - Y_{T_1+2T_2}\|_1] \leq 3n^{-A}$$

Finally, combining this with Markov's inequality proves Lemma 6.

## 8. COUPLING FOR NARROW MATRICES

In this section, we show that subset couplings are likely to succeed, and finish the proof of Theorem 2. The main lemma is:

**Lemma 17** (Coupling for Nearby Points). *Fix  $a > b+3$  and  $b, c > 0$ . Let  $X_t, Y_t$  be two copies of the chain constructed as above, so that after a burn-in period of length  $T_1 = 7(b+7.5)n \log(n)$  during which  $X_t, Y_t$  evolve by proportional coupling we have  $\|X_{T_1} - Y_{T_1}\|_1 < n^{-a}$ , then for  $t > T_1 + (\frac{1}{2} + c)n \log(n)$  we have  $\|\mathcal{L}(X_t) - \mathcal{L}(Y_t)\|_{TV} \leq 4n^{b+3-a} + 2n^{4-b} + n^{-c}$*

Construct a partition process from time  $T_1$  to time  $T = T_1 + (\frac{1}{2} + c)n \log(n)$ . Our first step is to define subset couplings and show that if  $X_t$  and  $Y_t$  are very close to each other and not too close to certain hyperplanes, then any subset couplings are likely to succeed. To define a subset coupling of  $X_t$  and  $Y_t$ , fix the subset  $S$  of interest and common update variables  $i = i(t) \in S$  and  $j = j(t) \notin S$ . If  $\delta_t[i, j]\epsilon_t[i, j] \geq 0$ , then the coupling of  $\lambda_t^x$  and  $\lambda_t^y$  is exactly as described for the other walk immediately before Lemma 8. Otherwise, assume  $\delta_y[i, j] < 0$  and  $\epsilon_t[i, j] > 0$ , and choose  $\lambda_t^x$  from  $[0, 1]$  uniformly at random. Then set  $\lambda_t^y$  to be the number which satisfies  $w(X_{t+1}, S) = w(Y_{t+1}, S)$  if such a number exists and is in the interval  $[0, 1]$ . Just as with equation (5), the measure (with mass less than 1) on  $\lambda_t^y$  that this assignment defines minorizes the uniform distribution, and so leaves a remainder distribution analogous to that given in equation (6). If there is no value of  $\lambda_t^y$  in  $[0, 1]$  which would allow  $w(X_{t+1}, S) = w(Y_{t+1}, S)$ , then choose  $\lambda_t^y$  uniformly from this remainder distribution. If  $\delta_y[i, j] > 0$  and  $\epsilon_t[i, j] < 0$ , the same construction works, but with  $\lambda_t^y$  chosen first, and  $\lambda_t^x$  chosen to satisfy  $w(X_{t+1}, S) = w(Y_{t+1}, S)$ .

Let  $p(X, Y, i, j, S)$  be the probability that a subset coupling of  $X, Y$  associated with subset  $S$  works given that coordinates  $i, j$  are updated. The proof of the following lemma is nearly identical to the proof of Lemma 4 in [26]:

**Lemma 18** (Subset Coupling). *Fix  $a - 2 > b > 0$ . For a pair of matrices  $(x, y)$  satisfying  $\sup_{m,k} |x[m, k] - y[m, k]| \leq n^{-a}$  and  $\inf_{m,k} (x[m, k], y[m, k], 2 - x[m, k], 2 - y[m, k]) \geq n^{-b}$ , we have for all sufficiently large  $n$  that  $p(x, y, i, j, S) \geq 1 - 4n^{b+2-a}$  uniformly in  $S \subset [n]$  and pairs  $i \in S, j \notin S$ .*

Next, as in Lemma 9, note that after a successful subset coupling involving sets  $S$  and  $R$  at time  $t$ , we have  $\|X_{t+1} - Y_{t+1}\|_{1,S} \leq \|X_t - Y_t\|_{1,S \cup R}$ . Thus, if all subset couplings until time  $t$  have succeeded,

$$(14) \quad \|X_t - Y_t\|_{1,A} \leq \|X_{T_1} - Y_{T_1}\|_1$$

for all  $S \in P_t$ .

We are ready to prove Lemma 17. By inequality (14) and Lemma 18, the probability of a subset coupling failing at some time  $t$  given all previous subset couplings have succeeded is at most  $4n^{b+2-a} + 2n^{2-b}$ . Using a union bound over all at most  $n - 1$  subset couplings, and then applying Lemma 4, the probability of all subset couplings succeeding and the partition process satisfying  $P_0 = \{[n]\}$  is at least  $1 - 4n^{b+3-a} - 2n^{4-b} - n^{-c}$ . By Lemma 5, this is a lower bound on the probability that  $X_T = Y_T$ , and thus by Lemma 3 an upper bound on the distance of  $X_T$  to stationarity. This proves the lemma.

Next, we put together Lemmas 11 and 17. Using the constants in those lemmas, we set  $b = c + 4$ ,  $a = 2c + 7$  and  $A = 3c + 7$ , to find that

$$\|\mathcal{L}(X_T) - \mathcal{L}(Y_T)\|_{TV} \leq 13n^{-c}$$

which is the upper bound in Theorem 2. To prove the lower bound, let  $\tau$  be the (random) first time at which all  $2n$  coordinates have been updated. Then fix the starting position  $X_0$  of the Markov chain and let  $H_{i,j} = \{X \in \mathbb{R}^{2n} | X[i, j] = X_0[i, j]\}$  and set  $H = \cup_{i,j} H_{i,j}$ . Then  $P[X_t \in H] - U_n(H) \geq P[\tau > t]$ . Since only four of  $2n$  coordinates are chosen at a time, the classical coupon-collector results in [12] tell us that at time  $T = \frac{1}{2}n(\log(n) - c)$ ,  $|K_n^T(x, H) - \pi(H)| \geq 1 - \exp(-\exp(c)) + o(1)$  as  $n$  goes to infinity.

## REFERENCES

- [1] David Aldous and Jim Fill. *Reversible Markov Chains and Random Walks on Graphs*. Unpublished Manuscript, 1994.
- [2] Olena Blumberg. A coupling proof for random transpositions. *Preprint*, 2011.
- [3] Bela Bollobas. *Random Graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Press Syndicate of the University of Cambridge, Cambridge, 2001.
- [4] R. Bubly and M. Dyer. Path coupling: a technique for proving rapid mixing in markov chains. *Symposium on Foundations of Computer Science*, pages 223–231, 1997.
- [5] Robert Burton and Yevgeniy Kovchegov. Mixing times via super-fast coupling. *Preprint*, 2011.
- [6] Persi Diaconis and Bradley Efron. Testing for independence in a two-way table: New interpretations of the chi-square statistic. *Ann. Stat.*, 13(3):845–874, 1985.
- [7] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, exponential families and orthogonal polynomials. *Statistical Science*, 23:151–178, 2008.
- [8] Persi Diaconis, Kshitij Khare, and Laurent Saloff-Coste. Gibbs sampling, conjugate priors and coupling. *Sankhya*, 72-A(1):136–169, 2009.
- [9] Persi Diaconis and Laurent Saloff-Coste. Comparison techniques for random walks on finite groups. *Annals of Probability*, 21(4):2131–2156, 1993.
- [10] Martin Dyer, Leslie Goldberg, Mark Jerrum, and Russel Martin. Markov chain comparison. *Probability Surveys*, 3:89–111, 2006.

- [11] Martin Dyer, Ravi Kannan, and John Mount. Sampling contingency tables. *Random Structures and Algorithms*, 10:487–506, 1997.
- [12] Paul Erdos and Alfred Renyi. On a classical problem of probability theory. *Magyar Tud. Akad. Mat. Kutato. Int. Kozl*, 1961.
- [13] Tom Hayes and Eric Vigoda. A non-markovian coupling for randomly sampling colorings. *FOCS proceedings*, 2003.
- [14] John Jiang. Polynomial mixing time of the kac random walk on the orthogonal group. *Preprint*, 2012.
- [15] David Levin, Yuval Peres, and Elizabeth Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, Rhode Island, 2009.
- [16] Jun Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. Springer, 2001.
- [17] Laslo Lovasz. Hit-and-run mixes fast. *Math. Prog.*, 1998.
- [18] Laslo Lovasz and Santosh Vempala. Hit and run is fast and fun. *Technical Report - Microsoft Research*, 2003.
- [19] Peter Matthews. Mixing rates for a random walk on the cube. *SIAM. J. on Algebraic and Discrete Methods*, 8(4):746–752, 1987.
- [20] Ben Morris. Random walks in convex sets. *PhD Thesis, University of California, Berkeley*, 2000.
- [21] Ben Morris. Improved bounds for sampling contingency tables. *Random Structures and Algorithms*, 21:135–146, 2002.
- [22] Dana Randall and Peter Winkler. Mixing points on a circle. *Lecture Notes in Computer Science*, 3624:426–435, 2005.
- [23] Dana Randall and Peter Winkler. Mixing points on an interval. *Proceedings of ANALCO*, 2005.
- [24] Jeffrey Rosenthal. Minorization conditions and convergence rates for markov chain monte carlo. *JASA*, 90:558–566, 1995.
- [25] Devarat Shah. Gossip algorithms. *Foundations and Trends in Networking*, 3(1):1–125, 2009.
- [26] Aaron Smith. A gibbs sampler on the n-simplex. *Preprint*, 2011.
- [27] Aaron Smith. Some analyses of markov chains by the coupling method. *PhD Thesis, Stanford University*, 2012.
- [28] Wai Kong Yuen. Applications of geometric bounds to convergence rates of of markov chains and markov processes on rn. *PhD Thesis, University of Toronto*, 2001.

DEPARTMENT OF MATHEMATICS, STANFORD UNIVERSITY, STANFORD, CA 94305  
*E-mail address:* `asmith3@math.stanford.edu`